NASA/TM—2018-219904

K-Means Cluster Study for Radiofrequency Propagation Characterization

Tyler Cody
University of Virginia, Charlottesville, Virginia

Rigoberto Roche and Janette C. Briones
Glenn Research Center, Cleveland, Ohio

July 2018

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program plays a key part in helping NASA maintain this important role.

The NASA STI Program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI Program provides access to the NASA Technical Report Server—Registered (NTRS Reg) and NASA Technical Report Server—Public (NTRS) thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counter-part of peer-reviewed formal professional papers, but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., "quick-release" reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question to help@sti.nasa.gov

- Fax your question to the NASA STI Information Desk at 757-864-6500

- Telephone the NASA STI Information Desk at 757-864-9658

- Write to:
  NASA STI Program
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM—2018-219904

K-Means Cluster Study for Radiofrequency Propagation Characterization

Tyler Cody
University of Virginia, Charlottesville, Virginia

Rigoberto Roche and Janette C. Briones
Glenn Research Center, Cleveland, Ohio

National Aeronautics and
Space Administration

Glenn Research Center
Cleveland, Ohio 44135

July 2018

Level of Review: This material has been technically reviewed by technical management.

Available from

NASA STI Program
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

This report is available in electronic form at http://www.sti.nasa.gov/ and http://ntrs.nasa.gov/

K-Means Cluster Study for Radiofrequency Propagation Characterization

Tyler Cody
University of Virginia
Charlottesville, Virginia 22903

Rigoberto Roche and Janette C. Briones
National Aeronautics and Space Administration
Glenn Research Center
Cleveland, Ohio 44135

Summary

The objective of this study was to design a simple method for mining radiofrequency (RF) propagation data. The study explored the characteristics of a large data set of propagation experiments conducted over the span of 5 years using several ground stations around the world. Furthermore, this study developed simple predictive models that can be used for link characterization and overall propagation behavior description, without the need for onsite physical measurements. It is understood that such statistical learning has several drawbacks in terms of accuracy and precision. K-means clustering was used to characterize the data set in a way never before explored in an attempt to create useful tools that reduce cost, time, and risk. Cosine distance was used as a method to determine the optimal number for clustering each feature. Dependence and independence analysis were performed to explore intrasensitivity and intersensitivity between the presented features, with respect to each other and time. Several predicative models were generated and evaluated with respect to a test set to assess a measure of prediction accuracy and precision. A simple method for data analysis was developed and tested as the basis for further studies and future refinement to produce optimal performing models.

Nomenclature

k	number of clusters used for K-means clustering
NaN	not a number
SCaN	Space Communications and Navigation
RF	radiofrequency

1.0 Background

The largest uncertainty in space communication system design lies in the impact of the atmospheric channel on propagating electromagnetic waves. The proper characterization of the atmosphere was needed to reduce cost and mitigate risk by ensuring the optimal design of the space-ground link. As NASA continues to work with Ka-band operations and optical frequencies, the understanding of this data and the development of predictive models has become crucial. The Space Communications and Navigation (SCaN) Ka-band Atmospheric Calibration/Radiofrequency (RF) Propagation Project involves a site characterization study, which was initiated to determine the atmospheric effects on Ka-band communication links at different sites, including Goldstone, California, White Sands, New Mexico, Guam, Canberra, Madrid, and Svalbard. The objective was to collect interferometer data of path length fluctuations and attenuation for 5 years at each site and maintain a database for use in Ka-band

propagation research and RF engineering to determine the optimized system link margin and system availability in the presence of atmospheric effects, such as rain, water vapor, and so forth. The goals of these measurements were to understand how propagation data can be used for defining requirements for a satellite communication system and to contribute to the development and improvement of models for the prediction of communication system performance.

NASA has been conducting extensive studies on the propagation of RF signals through the atmosphere when attempting to establish space links for communication with orbiting systems as well as other ground stations around the world. This project has produced a significant amount of data on the characteristics of the links each node has experienced, the overall propagation parameters of interest for the description of each event, and terms for optimal end-to-end communication. The data have been collected over the span of several years and from many ground stations around the world.

The primary purpose of the propagation project was to study how atmospheric effects influence the propagation of various forms of RF signals in order to characterize that distortion and come up with ways of mitigating it. This is done by conducting a number of experiments using different links, antennas, array configurations, and other methods. The characterization of such propagation is of the upmost importance to establish and maintain a high-fidelity link through the atmosphere. These characteristics also describe what conditions are favorable to such a link and, more importantly, what can be done to establish such a link, even if conditions are not favorable or they turn unfavorable as an event is occurring.

2.0 Introduction

The instrumentation of communication ground stations as well as the development of systems that work alongside such stations is one of the major cost items in the development and operation of such stations. The more instruments that are needed to characterize a specific link, the more cost associated with that particular station. There is also an added risk due to maintenance and the likelihood of systems being down during operations because of the intricacy and the sheer number of sensors necessary to measure all the desired phenomena affecting operations.

These ground stations operate as deterministic controlled systems. A controller turns the knobs either onsite or remotely. They generate timestamped data for all the operating sensors for each event as well as passive recording of atmospheric changes. Features such as power, wind speed, barometric pressure, and others are recorded at different frequencies of time, based on each sensor’s time tick and the configuration of each specific station.

A thorough understanding of atmospheric propagation is required to effectively design communication links. Therefore, the need to develop predictive models is critical to the design of efficient and effective communication systems. The objective of this study is to apply statistical learning techniques using the collected data to discern characteristics present in the data that may not have been otherwise explored or exploited until now. Furthermore, this study aims to produce a simple analysis method that can be applied with varied time resolutions to generate predictive models. These models can be used as a base for determining the expected characteristics of a link without the need for an array of sensors measuring every single parameter. A user can simply query the model and get an expected value for whatever feature is of interest.

3.0 Methods

A number of methods and processes were employed to facilitate and conduct the analysis. They are sequential and as follows:

1. Data processing
2. K-means heuristic for empty clusters
3. Finding the optimal number of clusters, *k*
4. Finding the best features and the extrapolation performance metric
5. The methods used to test generalization

They are described in this order in the following subsections.

3.1 Data Processing

The data were retrieved, which contained ground station data from Goldstone, California, Guam, Italy, Scotland, and White Sands, New Mexico. The data collected by each station varied. The power and weather data were selected from each station for analysis. The data at each station were stored in daily batches, with varying file formats, including .txt, .csv, and .dat files. At all stations, the power and weather data were collected at 60 samples per minute and 1 sample per minute, respectively.

Dialogue with domain experts indicated that the power signal was the only feature of interest from the power data. Note that the term feature is used to represent a type of information collected or computed from the raw data when it was not available directly. The availability varied both from station to station as well as over time within stations. When processing the data, the power signal was taken directly from the raw data unless it was unavailable, in which case it was calculated.

The instrumentation of the weather stations between and within each ground station varied. The shared features between daily batches of weather samples were made to have the same units of measurement. Often times, conversion was unnecessary.

The data from each station were processed from the daily readings into annual batches. The power and weather data were aligned using their date and timestamps. In total, across all stations, there are 35 years of data. If all samples with null values are removed, thus downselecting to the 1 sample per minute rate of weather and isolating to the columns that match across all years at all stations, then 17 million samples with seven features (not including the timestamp) were left. The features include power signal, air temperature, relative humidity, pressure, wind speed, wind direction, and rainfall. Several preprocessing methods were employed to restructure the data. These include batching, resampling, not-a-number (NaN)-removal, and many more.

The analysis was performed on a single year from a single station. Goldstone, California was selected because it was in operation for many successive years. The year 2013 was selected because data were available from roughly the entire year. The data were divided into weekly batches and randomly sampled each week to 17,000 samples. Weeks with less than 17,000 samples were dropped. The final set contained 35 weeks with 17,000 samples of six features. Note that rainfall was dropped because of lack of variation. The samples were chosen in weekly batches because it was a sufficiently large enough sample size and the intraday samples were often highly correlated. Each week was then split into sets for training and testing.

Figure 1 shows a diagram of the general process.

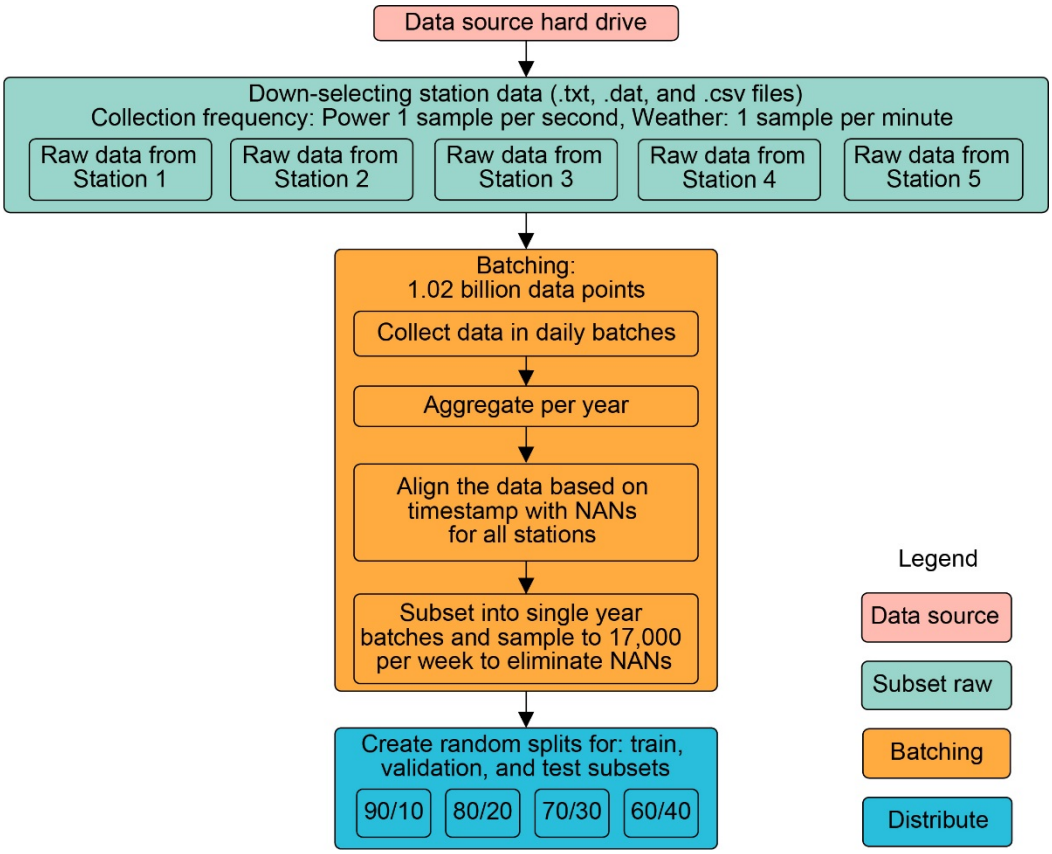

Figure 1.—Data ingestion and preprocess.

3.2 K-Means Heuristic for Empty Clusters

A variation on the K-means algorithm was used that employs a heuristic for dealing with empty clusters. During the training process after the labels are assigned to each point, the clusters are scanned for an empty cluster. If an empty cluster is found, the closest point from the largest cluster is added to the empty cluster. This is repeated until there are no longer empty clusters. Then, the centroid values are calculated as per the standard K-means algorithm and the learning process continues.

3.3 Finding Optimal *k* Value

A method for finding the optimal number of clusters, *k*, to use in the K-means algorithm was developed. Optimal is defined as the number of clusters when the marginal improvement in minimizing the cosine difference between the validation samples and their assigned centroids falls below an early stopping tolerance. The algorithm for finding the optimal *k* value is as follows:

1. Set the initial value of *k*, the maximum value of *k*, and the early stopping tolerance.
2. Initialize the previous cosine difference to a large value.
3. Start the training loop.
4. Run the K-means algorithm with *k* clusters on the training set.

5. Calculate the cosine difference between the validation samples and their assigned centroids.
   a. If the cosine difference is greater than the previous cosine difference, the absolute percent change between the current and previous cosine difference is less than the early stopping tolerance, or *k* equals the maximum value of *k*, then stop the training loop; *k* is at the optimal *k* value.
   b. If not, increase the *k* value and update the cosine difference.

   This algorithm is shown in Figure 2.

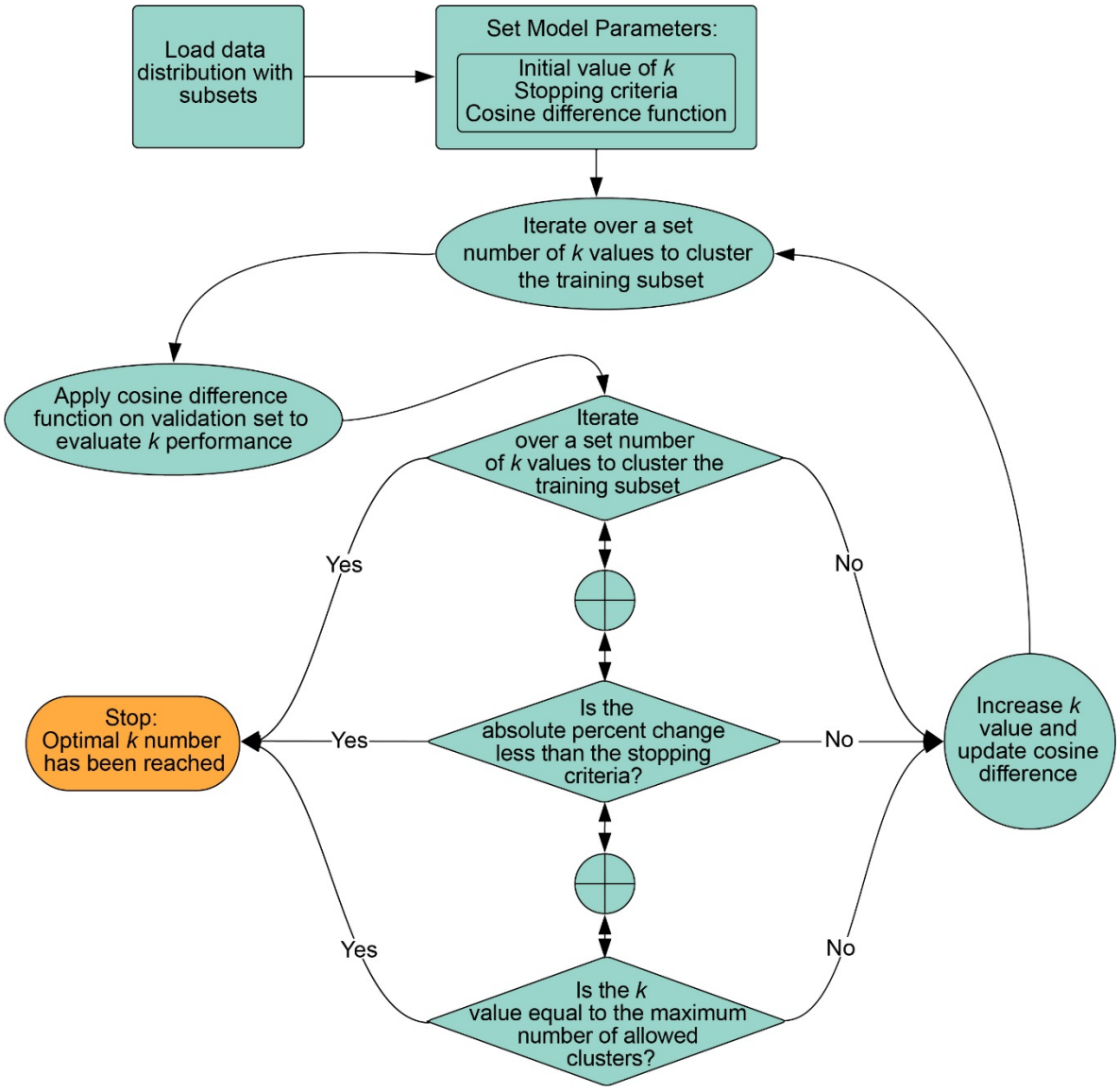

Figure 2.—Optimal *k* value algorithm.

3.4 Ranking Features With Extrapolation Performance Metric

An extrapolation performance metric was used to rank the features with respect to their ability to represent the entire data set. To establish this rank, first find the optimal number of clusters to use when clustering each feature individually. Note that the optimal number of clusters is found by monitoring the ability of the feature to cluster itself, not the data set as a whole. The set of optimal clusters was then used to calculate the extrapolation performance metric for each feature.

The extrapolation performance metric is determined as follows:

1. Group the training features into sets composed of samples assigned to the same cluster.
2. For each of these subsets, find the average value of each feature. For each cluster, the average value of the training samples assigned to the cluster for each feature not used to build the clusters has now been determined. The features not used to build the clusters are referred to as "unseen features."
3. Assign a label to each testing set cluster.
4. For each unseen feature, find the cosine difference between the testing samples and the averages corresponding to the feature and cluster label assigned to the samples. The extrapolation performance metric is the median of these cosine differences.

In short, the extrapolation performance metric is a cosine-difference-based measure of the ability of the feature to cluster the unseen features. Noting again that the clusters were not trained to optimize this ability to extrapolate, but rather the ability of the features to cluster themselves, it can be said that this extrapolation performance metric is a measure of the ability of the feature to represent the entire data set.

Specific to the analysis, there was concern with ranking this performance over weekly batches from a given year. The features were ranked by the upper bound of the 90-percent confidence interval on the performance metric. A 90-percent confidence gave a sufficiently high penalty to variance.

A dropoff in performance was used to select the best features. Specific to the analysis, relative humidity and wind direction were found to have noticeably higher performance than the other features. Air temperature, pressure, and wind speed had similar performance to each other, yet were worse than the top two, and power signal had the worst performance.

3.5 Generalization Test

An approach to test whether the ability to model dependent behavior increased the performance was taken to investigate the ability of the selected features to build clusters for extrapolating the value of unseen features. To do this, a fully independent and a fully dependent model were built and their performance was compared.

The first model, the independent model, is formed by training a set of clusters for each selected feature. The average value for each unseen feature at each cluster is found. To perform extrapolation on a given sample, the sample is split into its feature components. Each component is labeled by its respective clustering algorithm and assigned an average value for each unseen feature. The median of the components' average values for an unseen feature is the extrapolated value for that feature. In short, for each unseen feature, the model takes the median of the extrapolated values from independently trained clusters.

The second model, the dependent model, is formed by training a single set of clusters by using all of the selected features. The average value for each unseen feature at each cluster is found. In order to extrapolate the unseen features for a new sample, the cluster label for the sample is found. The

extrapolated values for the unseen features are the average values of the unseen features from the training sample at that cluster.

Comparing the performance of the independent and dependent models give insight into the benefit of incorporating dependence into model building.

4.0 Results

The process for ranking the features was carried out, and relative humidity and wind direction were selected as the best features. The generalization test was conducted using these features on each week, and overall, the independent model outperformed the dependent model. The following charts show the test set data in various representations for a single week out of the set of weeks used to conduct analysis. In all charts, the relative humidity is plotted in percent and the wind direction is plotted in degrees.

Figure 3 shows the density of the relative humidity and wind direction data. The darkness of a hexagonal region represents the number of points in that region. The relative humidity is mostly below 40 percent, except when the wind direction is above 200°, where the relative humidity varies more widely.

Figure 4 to Figure 7 are scatterplots and bubble plots of the clusters built for the independent model for a given week. Note that for the independent model, a set of clusters for wind direction and a set of clusters for relative humidity were built. The scatterplots show the data points, and the bubble charts show the size of the clusters and their location in the feature space. The scatterplots have random spacing applied so that the points do not all lie on a single line. Figure 4 and Figure 5 show the clusters for wind direction. Note that for wind direction, the values go from 0 to 360°, which wraps back to 0°. Figure 6 and Figure 7 confirm that a majority of the relative humidity readings are below 40 percent, and the largest clusters lie in this region.

For the independent model, because two separate sets of clusters are used, it is not easy to plot a representation of the independent model in a way that is intuitive to read, as can be done with the dependent model.

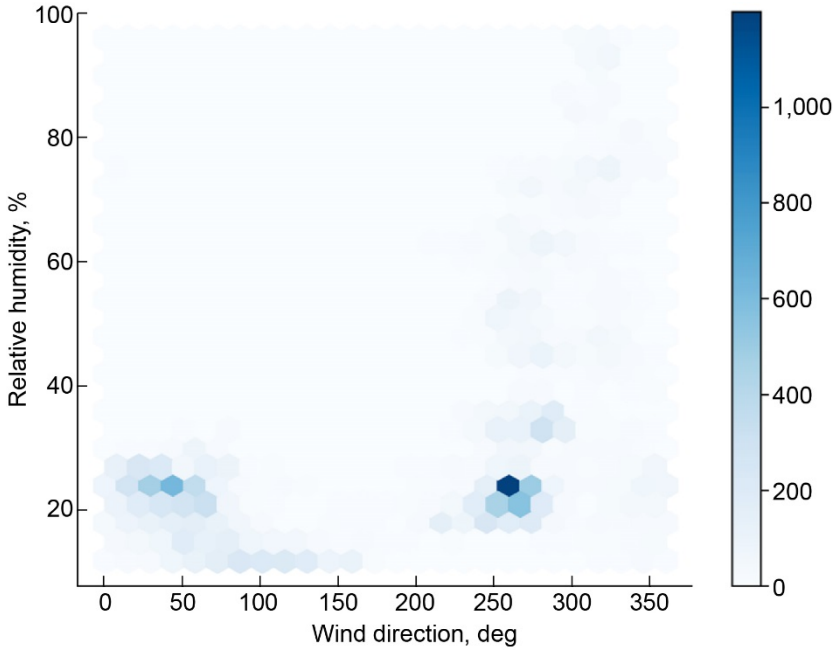

Figure 3.—Data density plot of relative humidity versus wind direction.

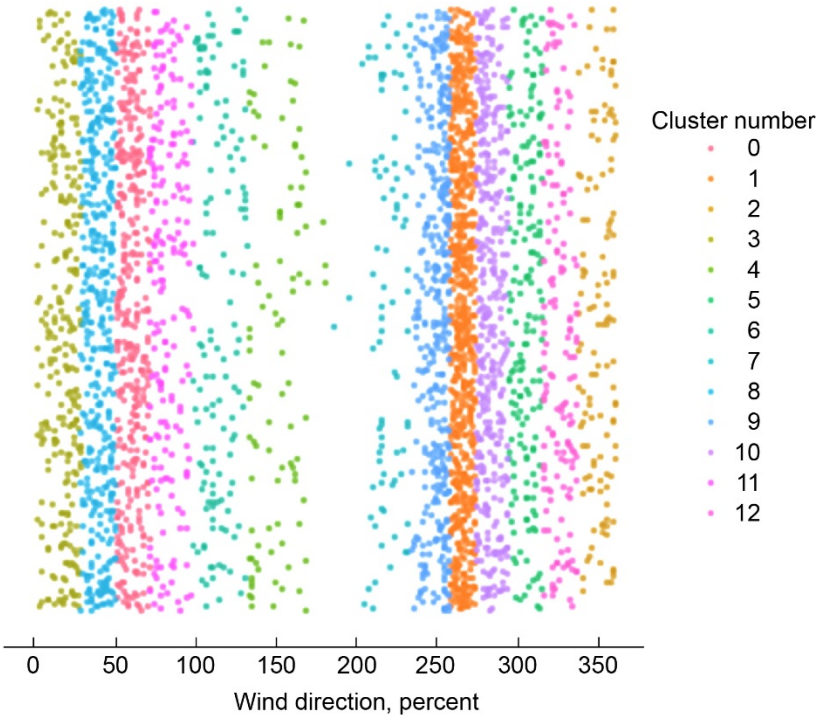

Figure 4.—Scatterplot of wind direction clusters for independent model.

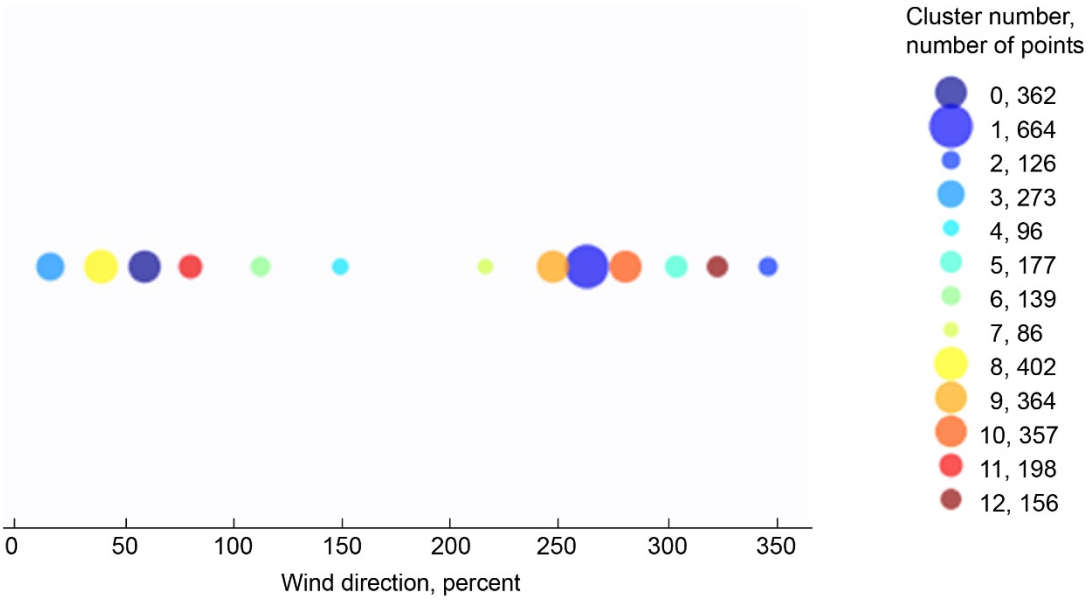

Figure 5.—Bubble plot of wind direction clusters for independent model.

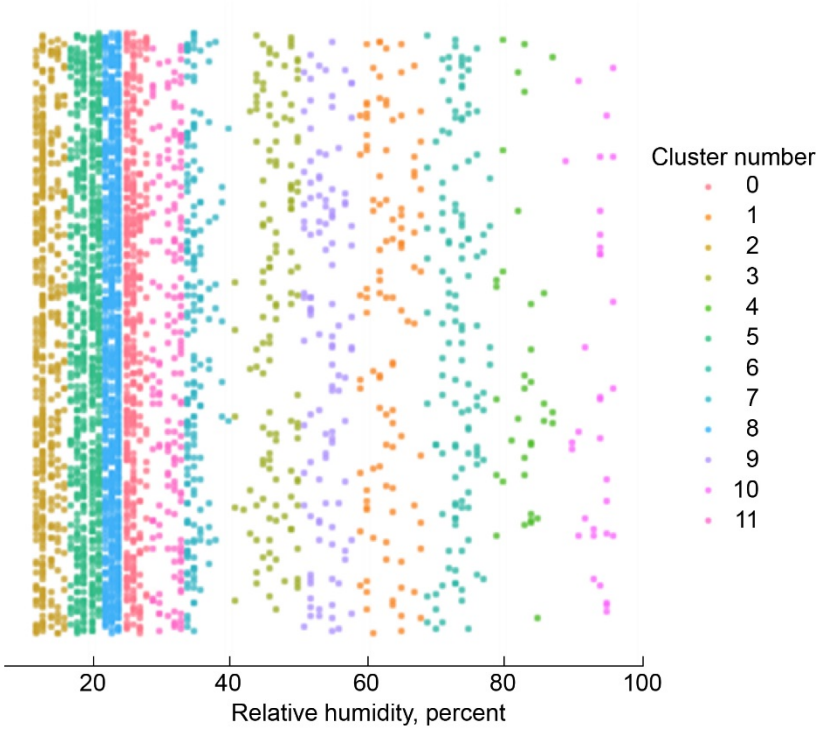

Figure 6.—Scatterplot of relative humidity clusters for independent model.

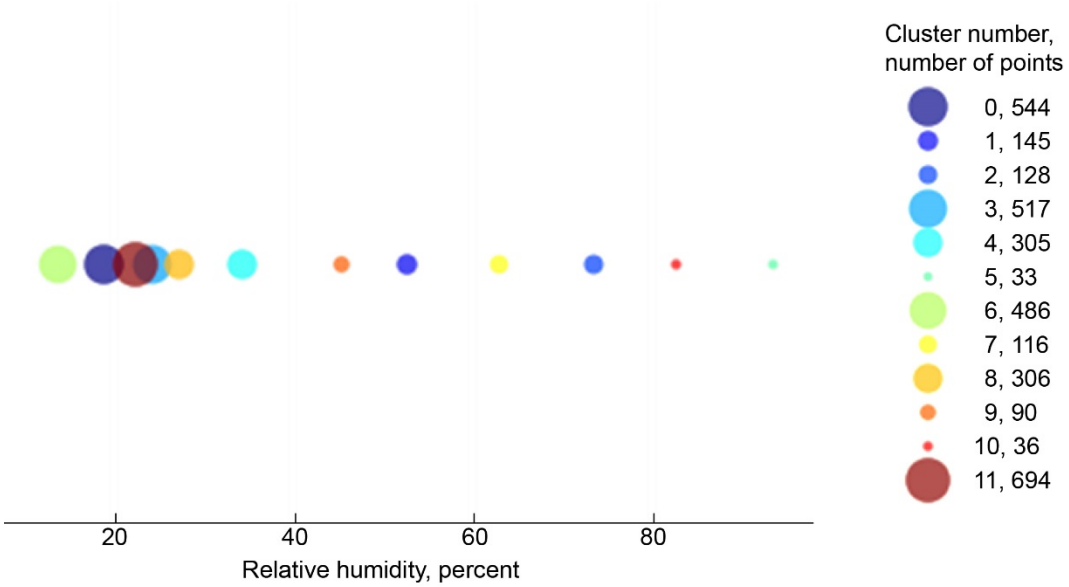

Figure 7.—Bubble plot of relative humidity clusters for independent model.

Figure 8 and Figure 9 show the scatterplot and bubble plot for the dependent model clusters. The dependent model consisted of one set of clusters built on all the best features, in this case, the wind direction and relative humidity. Figure 8 shows that the model was able to find clusters similar to how the data might be labeled manually. The number of clusters chosen by the algorithm seems appropriate under visual inspection.

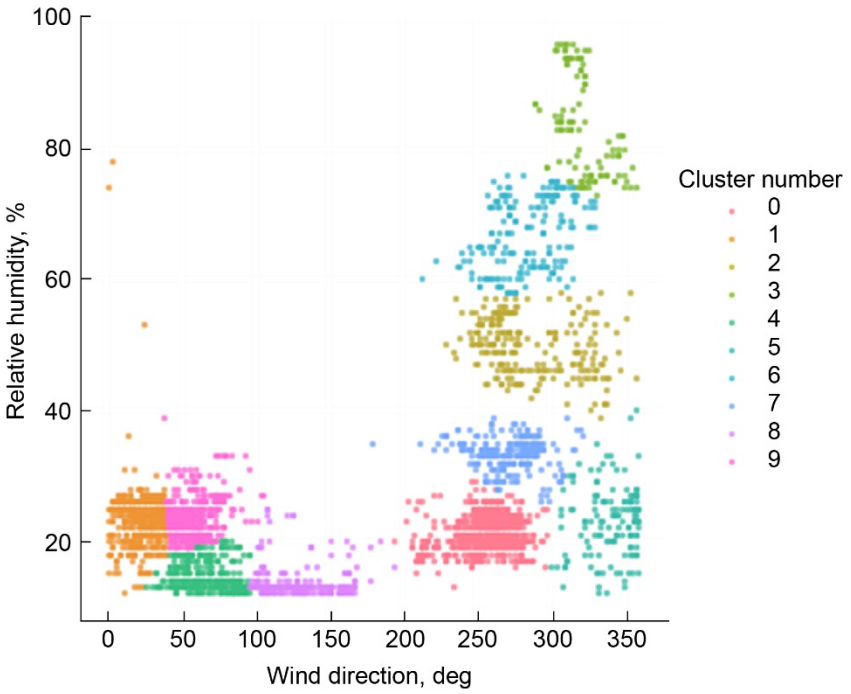

Figure 8.—Scatterplot of dependent model clusters (relative humidity vs. wind direction).

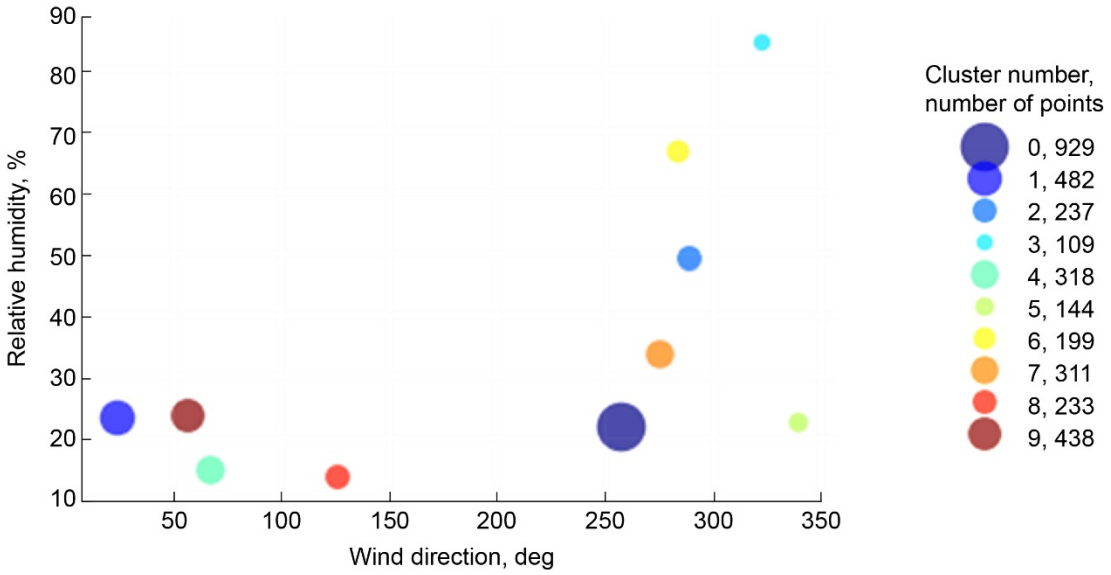

Figure 9.—Bubble plot of dependent model clusters (relative humidity vs. wind direction).

Figure 10 shows the decision boundaries of the dependent model. The sectors are the regions belonging to the cluster centers denoted with white triangles. The black points are data points from the test set. The center and top left portions of the chart show areas where the data has undefined behavior. This is evident by the large regions, the lack of clusters, and the lack of data points in those portions of the chart.

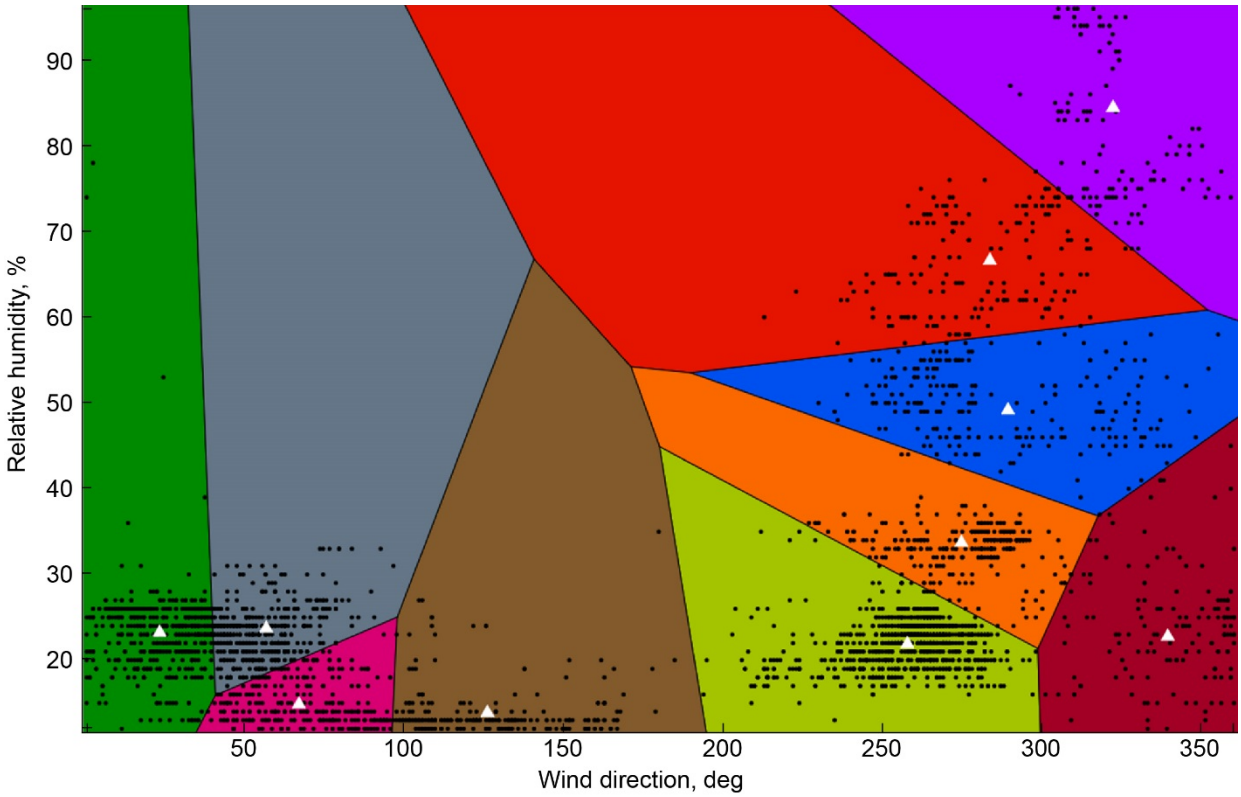

Figure 10.—Voronoi diagram of K-means decision boundaries for dependent model clusters.

These plots demonstrate the models and predictions produced by the methodology that was developed. The simple methods used were able to choose a seemingly reasonable *k* value for clustering the top features. This observation is based on visual inspection of the dependent model plots. The independent model plots cannot be as easily interpreted, however, they outperformed the dependent model in extrapolation and thus it may be reasonable to assume that the methods chose a similarly reasonable *k* value for clustering the features.

5.0 Concluding Remarks

In this study, a simple method for mining radiofrequency (RF) propagation data was described. The characteristics of a data set of propagation experiments were explored. Several simple predictive models were developed. These models were used for describing link characterization and overall propagation behavior. K-means clustering was used to analyze the data set by generating a number of centroids that characterize the features' bulk. Cosine distance was used as the loss function to determine the optimal number of *k* for clustering each feature. It was also used to evaluate the performance of predicted values versus measured values. Several clustering subsets were developed to test dependence and independence for each feature. This yielded a measure of intrasensitivity and intersensitivity between the presented features. The generated predictive models were evaluated with respect to a test set to assess accuracy and precision. The overall process for such methods was described in detail.

There are several benefits to applying such a processing method to this kind of data. One such benefit is the inherent simplicity of the model and how easily its analysis findings can be traced back to the mathematical implementation. Another benefit is the computational cost of performing such operations on such a large data set. This is ideal for a preliminary exploratory study of the data. This methodology

provides initial insight into what hidden patterns may be present and how to exploit them for benefiting link characterization research. Also, this analysis provides simple predicative models that can be reapplied to other data slices without having to change a single line of code.

There are some drawbacks to the approach presented in this study. The use of K-means clustering and cosine difference as methods for data mining bring inherent bias to the presented results. These methods have intrinsic mathematical limitations that can skew the expected values for feature prediction. For example, if another measure of precision was used to determine the optimal *k* number, say correlation, this may yield different results. However, for the purposes of data exploration, the simplest methods were chosen as an initial effort in describing such a data set. Another drawback to this analysis is the arbitrary time resolution chosen to generate the models. Although this was chosen in an effort to reduce computation load and explore several approaches, this resolution may not be optimal. This resolution may be too coarse to describe time-variant phenomena that the generated models are insensitive to. Further studies are necessary to determine whether increasing the time resolution to the order of hours or even minutes can yield better predictions. Note that such studies may require significant computation complexity and the generation of parallel processing applications to characterize such a large data set.

Bibliography

Pedregosa, Fabian, et al.: Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res., vol. 12, 2011, pp. 2825–2830.